

**Manuscript version: Author's Accepted Manuscript**

The version presented in WRAP is the author's accepted manuscript and may differ from the published version or Version of Record.

**Persistent WRAP URL:**

<http://wrap.warwick.ac.uk/117401>

**How to cite:**

Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

**Copyright and reuse:**

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

**Publisher's statement:**

Please refer to the repository item page, publisher's statement section, for further information.

For more information, please contact the WRAP Team at: [wrap@warwick.ac.uk](mailto:wrap@warwick.ac.uk).

# A framework for DNA quantification and outlier detection using multidimensional standard curves

Ahmad Moniri,<sup>†,§</sup> Jesus Rodriguez-Manzano,<sup>\*,†,§</sup> Kenny Malpartida-Cardenas,<sup>†</sup>  
Ling-Shan Yu,<sup>†</sup> Xavier Didelot,<sup>‡</sup> Alison Holmes,<sup>¶</sup> and Pantelis Georgiou<sup>†</sup>

<sup>†</sup>*Centre for Bio-Inspired Technology, Department of Electrical and Electronic Engineering,  
Imperial College London, London, UK*

<sup>‡</sup>*School of Life Sciences and Department of Statistics,  
University of Warwick, Coventry, UK*

<sup>¶</sup>*NIHR Health Protection Research Unit, Healthcare Associated Infections and  
Antimicrobial Resistance, Imperial College London, London, UK*

<sup>§</sup>*These authors contributed equally to this work as first authors.*

E-mail: j.rodriguez-manzano@imperial.ac.uk  
Phone: +44 207 5940843

## Abstract

Real-time PCR is a highly sensitive and powerful technology for the quantification of DNA and has become the method of choice in microbiology, bioengineering and molecular biology. Currently, the analysis of real-time PCR data is hampered by only considering a single feature of the amplification profile to generate a standard curve. The current “gold standard” is the cycle-threshold ( $C_t$ ) method which is known to provide poor quantification under inconsistent reaction efficiencies. Multiple single-feature methods have been developed to overcome the limitations of the  $C_t$  method, however, there is an unexplored area of combining multiple features in order to benefit from their joint information. Here, we propose a novel framework that combines existing standard curve methods into a *multidimensional standard curve*. This is achieved by considering multiple features together such that each amplification curve is viewed as a point in a multidimensional space. Contrary to only consid-

ering a single-feature, in the multidimensional space, data points do not fall exactly on the standard curve, which enables a similarity measure between amplification curves based on distances between data points. We show that this framework expands the capabilities of standard curves in order to: optimise quantification performance, provide a measure of how suitable an amplification curve is for a standard, and thus automatically detect outliers and increase the reliability of quantification. Our aim is to provide an affordable solution to enhance existing diagnostic settings through maximizing the amount of information extracted from conventional instruments.

## Introduction

The real-time polymerase chain reaction (qPCR) has become a routine technique in microbiology, bioengineering and molecular biology for detecting and quantifying nucleic acids.<sup>1–3</sup> This is predominantly due to its large

dynamic range (7-8 magnitudes), desirable sensitivity (5-10 molecules per reaction) and reproducible quantification results.<sup>4-6</sup> New methods to improve the analysis of qPCR data are invaluable to a number of application fields, including environmental monitoring and clinical diagnostics.<sup>7-10</sup>

The current "gold standard" for absolute quantification of DNA (or RNA if preceded by a reverse transcription step) using standard curves is the cycle-threshold ( $C_t$ ) method.<sup>11-13</sup> The  $C_t$  value is a feature of the amplification curve defined as the cycle number in the exponential region from which there is a detectable increase in fluorescence. However, this method is known to provide inaccurate quantification under inconsistent reaction efficiencies.

Since the  $C_t$  method was proposed, several alternative methods have been developed to improve absolute quantification in terms of accuracy, precision and robustness. The focus of current research is based on the computation of single features, for example,  $Cy_0$  or  $-\log_{10}(F_0)$ , that are linearly related to initial concentration, as in  $C_t$ .<sup>14,15</sup> Each feature corresponds to an underlying assumption, for example, the  $C_t$  approach assumes the PCR efficiency to be constant between reactions and cycles.<sup>11</sup> The  $Cy_0$  approach allows for different efficiency between reactions but assumes a constant efficiency between cycles.<sup>14</sup> The third feature,  $-\log_{10}(F_0)$ , allows for different efficiency between reactions but additionally assumes that it decreases from cycle to cycle.<sup>15</sup> These single-feature methods provide a simple approach for absolute quantification, however, the degrees of freedom to implement more complex data analysis techniques is limited, and the use of multiple features together has been unexplored.

Inspired by the field of Machine Learning, this paper takes a multidimensional view, combining multiple features in order to take advantage of the information and principles behind all of the current standard curve methods developed. Here, we provide a novel framework that combines existing standard curve methods into a *multidimensional standard curve* (MSC). This is achieved by considering multiple features together such that each amplifica-

tion curve is viewed as a point in a multidimensional space. Therefore, the standard curve in the multidimensional space should theoretically form a 1D line. Contrary to only considering a single-feature, in the multidimensional space, data points do not fall exactly on the standard curve and thus enables a similarity measure between amplification curves based on distances between data points.

We show that this framework expands the capabilities of standard curves in order to: optimise quantification performance, provide a measure of how suitable an amplification curve is for a standard, and thus automatically detect outliers and increase the reliability of quantification. Here, *outlier* refers to abnormal amplification data, due to non-specific target amplification or inconsistencies in amplification efficiency and reaction conditions (e.g. annealing temperature).

This has been demonstrated through constructing an MSC for phage lambda DNA and evaluating the quantification performance using a figure of merit combining accuracy, precision and overall predictive power. Following this, we evaluated the framework for outlier detection using non-specific DNA targets where we explored the notion of distance in the multidimensional space to understand if it can be used as a similarity measure between amplification curves. Finally, we used annealing temperature variation as a proxy for amplification efficiency in order to investigate whether the MSC can be used to disregard specific outliers and enhance quantification.

In Rodriguez-Manzano et al., it was shown that, using the MSC methodology described in the present manuscript, it is possible to simultaneously perform single-channel quantification and multiplexing of the four most prominent carbapenem-resistant genes.<sup>16</sup> We hope that by sharing this framework, others will be able to adapt and build upon this work to meet their objectives and explore new capabilities enabled by MSCs.

# Experimental Section

## Proposed Framework

In order to understand the proposed framework, it is useful to have an overall picture of how standard curves are used for quantification. Here, two terms, namely *training* and *testing* are borrowed from Machine Learning to describe the construction of a standard curve and quantify unknown samples respectively. Within the conventional single-feature approach, training is typically achieved through 4 stages: pre-processing, curve fitting, feature extraction and line fitting (linear regression). This is illustrated in Fig 1 (top branch and solid line). Testing is accomplished by extracting the same feature as when training, and using the generated standard curve to quantify the concentration in unknown samples.

The proposed framework extends the conventional approach by increasing the dimensionality of the standard curve in order to explore and take advantage of using multiple features together. This new framework is presented in Fig 1 (bottom branch and dotted line). For training, there are 6 stages: pre-processing, curve fitting, multi-feature extraction, high dimensional line fitting, multidimensional analysis and dimensionality reduction. Within this framework, testing can be achieved through dimensionality reduction, and multidimensional analysis using the MSC can be used to detect outliers and support quantification.

In contrast with conventional training, instead of extracting a single linear feature, multiple features are extracted from the processed amplification curves, for example, denoted using the dummy labels X, Y and Z. Therefore, each amplification curve has been reduced to 3 values (e.g.  $X_1$ ,  $Y_1$  and  $Z_1$ ) and, consequently, can be viewed as a point in 3 dimensional space. *It is important to stress that any number of features could be used as long as they are linearly related to the initial target concentration.* Therefore, the training data should theoretically form a 1-D line in 3-D space. This line is approximated using high-dimensional line fitting and generates what is called the *multidimensional standard curve*. Although, the train-

ing data forms a line, it is important to understand that data points do not lie exactly on the line. Consequently, there is considerable room for exploring this multidimensional space, referred to as the *feature space*, which will be also reported in this paper.

For quantification purposes, the MSC needs to be mapped into a single dimension, denoted as  $M_0$ , linearly related to the initial concentration of the target. In order to distinguish this curve from conventional standard curves, it is referred to here as the *quantification curve*. This can be achieved using dimensionality reduction techniques (DRT).<sup>17</sup> Mathematically, this means that DRTs are multivariate functions of the form:  $M_0 = \phi(X, Y, Z)$  where  $\phi(\cdot) : \mathcal{R}^3 \rightarrow \mathcal{R}$ . In fact, given that scaling features does not affect linearity,  $M_0$  can be mathematically expressed as  $M_0 = \phi(\alpha_1 X, \alpha_2 Y, \alpha_3 Z)$  where  $\alpha_i$  for  $i \in \{1, 2, 3\}$  are scalar constants. These weights provide a simple method for choosing the contribution of each individual feature in order to improve quantification. Furthermore, regardless of the weightings, all features will be considered for the multidimensional analysis.

## Multidimensional Standard Curves

In this section, we provide the specific *instance of framework* used to construct the MSC in this study.

i) *Preprocessing*. The first step in data analysis is to perform background subtraction. This is accomplished by subtracting the average of the fluorescent readings in the first five cycles from every amplification curve. More advanced methods could be considered to improve performance such as the *taking-difference linear regression method*.<sup>18</sup>

ii) *Curve Fitting*. In this study, the model used to fit the amplification curves is the 5-parameter sigmoid (Richards Curve) given by:

$$F(x) = F_b + \frac{F_{\max}}{(1 + e^{-(x-c)/b})^d} \quad (1)$$

where  $x$  is the cycle number,  $F(x)$  is the fluorescence at cycle  $x$ ,  $F_b$  is the background fluorescence,  $F_{\max}$  is the maximum fluorescence,  $c$

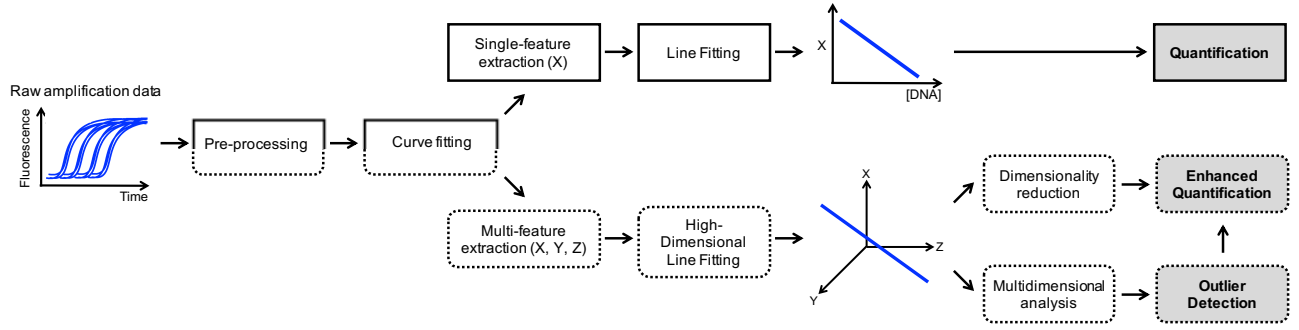


Figure 1: Block diagram showing the conventional method (top branch and solid line) compared to the proposed framework (bottom branch and dotted line) for target quantification. In both cases, raw amplification data for several known concentrations of the target are typically pre-processed and fitted with an appropriate curve. In the conventional case, a single feature such as the cycle threshold,  $C_t$ , is extracted from each curve. Subsequently, the extracted features are graphed as a function of concentration and a line is fit to the data in order to generate a standard curve and quantify unknown samples. In the proposed framework, multiple features are extracted and thus a 1D line in high dimensional space (called the feature space) is fitted in order to construct a multidimensional standard curve. Through dimensionality reduction, enhanced quantification can be achieved and performing multidimensional analysis in the feature space allows for outlier detection. The quantification can be further assisted by disregarding outliers.

is the fractional cycle of the inflection point,  $b$  is related to the slope of the curve and  $d$  allows for an asymmetric shape (Richard’s coefficient). The optimisation algorithm used to fit the five parameters of this model to the data is the trust-region method and is based on the interior-reflective Newton method.<sup>19,20</sup> Here, the trust-region method is chosen over the Levenberg-Marquardt algorithm since bounds for the 5 parameters can be chosen in order to encourage a unique and realistic solution.<sup>21,22</sup> The lower and upper bounds for the 5 parameters,  $[F_b, F_{\max}, c, b, d]$ , are given as:  $[-0.5, -0.5, 0, 0, 0.7]$  and  $[0.5, 0.5, 50, 100, 10]$ , respectively.

iii) *Feature Extraction.* Three features were used to construct the multidimensional standard curve in this study:  $C_t$ ,  $Cy_0$  and  $-\log_{10}(F_0)$ . Therefore, each amplification curve can be represented as a point in 3-dimensional space, i.e.  $\mathbf{p} = [C_t, Cy_0, -\log_{10}(F_0)]^T$  where  $[\cdot]^T$  denotes the transpose operator. Note that by convention, for the formulas in this paper, vectors are denoted using bold lowercase letters.

The cycle-threshold,  $C_t$ , is computed by fitting the amplification curve with the 5-parameter sigmoid in equation 1, normalizing the fitted function  $F(x)$  with respect to  $F_{\max}$ , and then takes  $C_t$  as the time where  $F(x)$

exceeds 0.2 (i.e. 20% of its maximum fluorescence). The  $Cy_0$  approach, proposed by Guescini et al., also uses the 5-parameter sigmoidal curve-fitting and takes  $Cy_0$  as the intersection between the abscissa axis and the tangent of the inflection point in the fitted  $F(x)$ .<sup>14</sup> The third feature,  $F_0$ , proposed by Rutledge, fits the sigmoid up to a “cut-off cycle” and takes  $F_0$  as the fluorescence at cycle 0.<sup>15</sup> Note that,  $-\log_{10}(F_0)$  is used instead of  $F_0$  since it is linearly related to initial template concentration.

iv) *Line Fitting.* In this work, to fit a 1D line to the training data in multidimensional space, i.e. construct the MSC, the method of least squares is used. Or, equivalently, by using the first principal direction in principal component analysis (PCA).<sup>23</sup> If sufficient data exists, other methods such as random sample consensus (RANSAC) which are robust to outliers could be used.<sup>24</sup>

v) *Similarity Measure.* There are two similarity measures used in this study: Euclidean and Mahalanobis distance. The Euclidean distance between a point,  $\mathbf{p}$ , and the MSC can be calculated by orthogonally projecting the point onto the MSC and then using simple geometry

to calculate the Euclidean distance,  $e$ , given by

$$P = \Phi(\mathbf{p}, \mathbf{q}_1, \mathbf{q}_2) = \frac{(\mathbf{p} - \mathbf{q}_1)^T(\mathbf{q}_2 - \mathbf{q}_1)}{(\mathbf{q}_2 - \mathbf{q}_1)^T(\mathbf{q}_2 - \mathbf{q}_1)} \quad (2)$$

$$e = |(\mathbf{p} - \mathbf{q}_1) - (\mathbf{q}_1 + P \cdot (\mathbf{q}_2 - \mathbf{q}_1))| \quad (3)$$

where  $\Phi$  computes the projection of the point  $\mathbf{p} \in \mathcal{R}^n$  onto the multidimensional standard curve, the points  $\mathbf{q}_1, \mathbf{q}_2 \in \mathcal{R}^n$  are any two distinct points that lie on the standard curve, and  $|\cdot|$  denotes the absolute value operator.

The Mahalanobis distance is defined as the distance between a point,  $\mathbf{p}$ , and a distribution,  $\mathcal{D}$ , in multidimensional space.<sup>25</sup> Similar to the Euclidean distance, the point is first projected onto the MSC and the following formula is applied to compute the Mahalanobis distance,  $d$ ,

$$d = \sqrt{(\mathbf{p} - P(\mathbf{q}_2 - \mathbf{q}_1))^T \Sigma^{-1} (\mathbf{p} - P(\mathbf{q}_2 - \mathbf{q}_1))} \quad (4)$$

where  $\mathbf{p}$ ,  $P$ ,  $\mathbf{q}_1$  and  $\mathbf{q}_2$  are given in equation (2) and  $\Sigma$  is the co-variance matrix of the training data used to approximate the distribution  $\mathcal{D}$ .

It can be shown that if the data is approximately normally distributed then the Mahalanobis distance squared,  $d^2$ , follows a  $\chi^2$ -distribution.<sup>26</sup> Therefore, a  $\chi^2$ -distribution table can be used to translate a specific  $p$ -value into a distance threshold. For instance, for a  $\chi^2$ -distribution with 2 degrees of freedom, a  $p$ -value of 0.001 corresponds to a Mahalanobis distance of 3.72.

vi) *Feature Weights*. As mentioned previously, in order to maximize quantification performance, different weights,  $\alpha$ , can be assigned to each feature. This can be accomplished by minimizing an error measure on the training data, where quantities of template are known, using an optimization algorithm. The specific error measure used in this study is described in the following subsection. The optimisation algorithm is the Nelder-Mead simplex algorithm with weights initialised to unity, i.e. beginning with no assumption on the quantification performance of individual features.<sup>27,28</sup> This is a standard algorithm and only 20 iterations are used to find the weights so that there is little computational overhead.

vii) *Dimensionality Reduction*. In this study, every point of the MSC is mapped into an estimated concentration using principal component regression, i.e.  $M_0 = P$  from equation (2). This is compared with projecting the MSC onto all three dimensions, i.e.  $C_t$ ,  $C_{y_0}$  and  $-\log_{10}(F_0)$ .

## Evaluating Standard Curves

In consistency with the current literature on evaluating standard curves, relative error (RE) and coefficient of variation (CV) are used to measure accuracy and precision respectively. The CV for each concentration is calculated after normalising the standard curves such that a fair comparison across standard curves is achieved. The formula for RE is given by:

$$\text{RE}_i = 100 \times \left| \frac{\hat{x}_i}{x_i} - 1 \right| \quad (5)$$

where  $i$  is the index of a given training point,  $x_i$  is the true template concentration of the  $i^{\text{th}}$  training data, and  $\hat{x}_i$  is the estimate of  $x_i$  using the standard curve. The CV for a given concentration is computed as:

$$\text{CV}_i = 100 \times \frac{\text{std}(\hat{\mathbf{x}}^i)}{\text{mean}(\hat{\mathbf{x}}^i)} \quad (6)$$

where  $i$  is the index of a given training point and  $\hat{\mathbf{x}}^j$  is a vector of estimated concentrations for all training points with the same concentration as  $x_i$ . The sample standard deviation and sample mean are denoted by  $\text{std}(\cdot)$  and  $\text{mean}(\cdot)$ , respectively. This paper also uses the leave-one-out cross validation (LOOCV) error as a measure for stability and overall predictive performance.<sup>23</sup> Stability refers to the predictive performance when points are removed from the training process. The LOOCV is given as:

$$\text{LOOCV}_i = 100 \times \left| \frac{\hat{z}_i}{x_i} - 1 \right| \quad (7)$$

where  $i$  is the index of a given training point,  $x_i$  is the true concentration of the  $i^{\text{th}}$  training data and  $\hat{z}_i$  is the estimate of  $x_i$  using a standard curve generated without the  $i^{\text{th}}$  training point. In this study, the LOOCV is specified as a percentage in order to compare across differ-

ent template concentrations, as shown in equation 7.

In order for the optimisation algorithm to compute  $\alpha$  and simultaneously minimise the three aforementioned measures, it is convenient to introduce a figure of merit,  $Q$ , to capture all of the desired properties. For a given training point, the product between all three errors,  $Q_i$ , can be used to heuristically measure the quantification performance. Therefore,  $Q$  can be defined as the average over all  $Q_i$ , as shown in equation 8, and is the error measure that the optimisation algorithm will minimise.

$$Q = \frac{1}{N} \sum_{i=1}^N \text{RE}_i \times \text{CV}_i \times \text{LOOCV}_i \quad (8)$$

## Statistical Analysis

The  $p$ -values used for assessing the significance between methods in absolute quantification were calculated using a paired, two-sided Wilcoxon signed rank test. Statistical significance was considered as  $*p\text{-value} < 0.05$ ,  $**p\text{-value} < 0.01$ ,  $***p\text{-value} < 0.001$ ,  $****p\text{-value} < 0.0001$ . Outliers using multidimensional standard curves were determined using a  $\chi^2$ -distribution with 2 degrees of freedom and statistical significance was assumed for a  $p\text{-value} < 0.001$ . The Henze-Zirkler test is used to determine multivariate normality with a  $p\text{-value}$  significance level of 0.05.<sup>29</sup>

## Fluorescence Datasets

DNA targets used for qPCR experiments in this study:

i) Standard curves were constructed using synthetic double-stranded DNA (gblocks Fragments Genes) containing phage lambda DNA sequence (DNA concentration ranging from  $10^2$  to  $10^8$  copies per reaction). Reactions were performed at an annealing temperature of  $62^\circ\text{C}$ . See Table S1 for primer and sequence information.

ii) Non-specific outlier detection experiments were performed using synthetic double-stranded DNA carrying *bla*<sub>OXA-48</sub>, *bla*<sub>NDM</sub> and

*bla*<sub>KPC</sub> genes, in this work referred to as outlier 1, 2 and 3 respectively. Reactions were performed at annealing temperature of  $68^\circ\text{C}$ . See Table S2-S4 for primer and sequence information.

iii) Specific outlier detection experiments were performed using synthetic double-stranded DNA containing lambda DNA sequence at  $10^5$  copies per reaction. Reactions were performed at annealing temperatures ranging from  $54.0$  to  $73.6^\circ\text{C}$ . See Table S1 for primer and sequence information.

All oligonucleotides were synthesised by IDT (Integrated DNA Technologies, Germany) with no additional purification. The specific PCR primers for lambda phage were designed in-house using Primer3 ([http://biotoools.umassmed.edu/bioapps/primer3\\_www.cgi](http://biotoools.umassmed.edu/bioapps/primer3_www.cgi)), whereas the primer pairs used for the outlier detection were taken from Monteiro et al.<sup>30</sup> Real-time PCR amplifications were conducted using FastStart Essential DNA Green Master kit (Roche Diagnostics, Germany) according to manufacturer's instructions. Each reaction consisted of  $2.5 \mu\text{L}$  FastStart Essential DNA Green Master  $2\times$  concentrated,  $1 \mu\text{L}$  of PCR grade water,  $0.5 \mu\text{L}$  of  $10\times$  primer mixture at  $5\mu\text{M}$  and  $1 \mu\text{L}$  of DNA at variable amounts, in a  $5 \mu\text{L}$  final reaction volume. Thermocycling was performed using a LightCycler 96 (Roche) initiated by a 10 min incubation at  $95^\circ\text{C}$ , followed by 40 cycles:  $95^\circ\text{C}$  for 20 sec;  $62^\circ\text{C}$  (for lambda) or  $68^\circ\text{C}$  (for non-specific outliers) for 45 sec; and  $72^\circ\text{C}$  for 30 sec, with a single fluorescence reading taken at the end of each cycle. Each reaction combination was conducted in quintuplicates/octuplicates. All the runs were completed with a melting curve analysis performed at  $95^\circ\text{C}$  for 10 sec,  $65^\circ\text{C}$  for 60 sec, and  $97^\circ\text{C}$  for 1 sec (continuous reading from  $65$  to  $97^\circ\text{C}$ ) to confirm the specificity of amplification and lack of primer dimer. Appropriate positive and negative controls were included in each experiment.

# Results and Discussion

In this study, a new framework is presented to construct multidimensional standard curves in order to: (i) optimise the quantification performance; (ii) detect outliers; and (iii) provide a heuristic measure for the similarity between an amplification curve and the MSC.

**Optimising Quantification Performance.** Synthetic phage lambda DNA was used to construct an MSC and evaluate its quantification performance relative to single feature methods. The resulting MSC, constructed using the features  $C_t$ ,  $C_{y_0}$  and  $-\log_{10}(F_0)$ , is visualised in Fig 2 (a). The computed features and curve-fitting parameters for each amplification curve grouped by concentration, ranging from  $10^2$  to  $10^8$  copies per reaction, is presented in Table S5. For comparison, Fig 2 (b) shows the quantification curves for each single-feature method plus the multi-feature method  $M_0$  which is obtained after dimensionality reduction through principal component regression.

The proposed framework enables the user to optimise quantification performance (through weighting each feature) in terms of a figure of merit,  $Q$ . The  $Q$  chosen in this work combines RE, CV and LOOCV. After 20 iterations of the optimisation algorithm, the weights  $\alpha$  converged to  $[-0.0741, 1.1185, 1.6574]$  corresponding to  $C_t$ ,  $C_{y_0}$  and  $-\log_{10}(F_0)$  respectively. It is important to stress that although the optimization algorithm suggests different performance across the selected features, there is value in keeping all of them as it can assist outlier detection, as shown in the subsequent sections.

The average  $Q$  ( $\pm$  standard deviation) for  $M_0$  against the single-feature methods is visualised in Fig 2 (c), where for  $C_t$ ,  $C_{y_0}$ ,  $-\log_{10}(F_0)$  and  $M_0$  it is  $4587 \pm 12799$ ,  $3327 \pm 10357$ ,  $13384 \pm 19966$  and  $2547 \pm 8058$  respectively. Therefore, in terms of average  $Q$ ,  $M_0$  enhances quantification by 17.44% ( $p < 0.05$ ), 10.65% ( $p < 0.05$ ) and 99.3% ( $p < 0.0001$ ) compared to  $C_t$ ,  $C_{y_0}$  and  $-\log_{10}(F_0)$  respectively. A summary and breakdown of each calculated error for all methods grouped by concentration are provided in Table S6-S9.

**Outlier Detection.** In this section, the concept of distance in the feature space is explored in order to demonstrate the capability of the framework for outlier detection. The term *outlier* refers to abnormal amplification curves with respect to the lambda DNA training data. This can be caused by non-specific amplification or inconsistent amplification efficiencies and reaction conditions; both of which are investigated in this report.

First, three non-specific DNA targets with respect to the phage lambda MSC, referred to as outlier 1, 2 and 3, were amplified. Subsequently, features were extracted from each amplification curve following the same procedure as the training data for the phage lambda MSC. Finally, the outliers were plotted in the feature space, as shown in Fig 3. It is visually clear that the outliers do not lie on the MSC and, therefore, this suggests that sufficient information is captured from the three features extracted from the amplification curves in order to distinguish the different targets from phage lambda DNA. Notice that without any secondary confirmation, e.g. from melting curves or agarose gels, the test data itself suggests it is not ‘similar’ to the training data.

In order to fully capture the position of the outliers in the feature space, it is convenient to remove the effect of concentration and view the feature space along the axis of the multidimensional standard curve. This is achieved by projecting all the data points in the feature space onto the plane perpendicular to the standard curve as illustrated in Fig 4 (a). The resulting projected points are shown in Fig 4 (b). It is clear that all three outliers can be clustered and clearly distinguished from the training data. Furthermore, the Euclidean distance,  $e$ , from the MSC to the mean of the outliers is given by  $e_1 = 1.44$ ,  $e_2 = 0.99$  and  $e_3 = 1.66$ . Given that the furthest training point from the MSC in terms of Euclidean distance is 0.36, the ratio between  $e_1$ ,  $e_2$ ,  $e_3$  and the furthest training point is 4.00, 2.75 and 4.61, respectively. In other words, the mean of outlier 1 is 4 times further than the furthest training point. Therefore, this ratio can be used as a similarity measure and the three clusters could be classified as



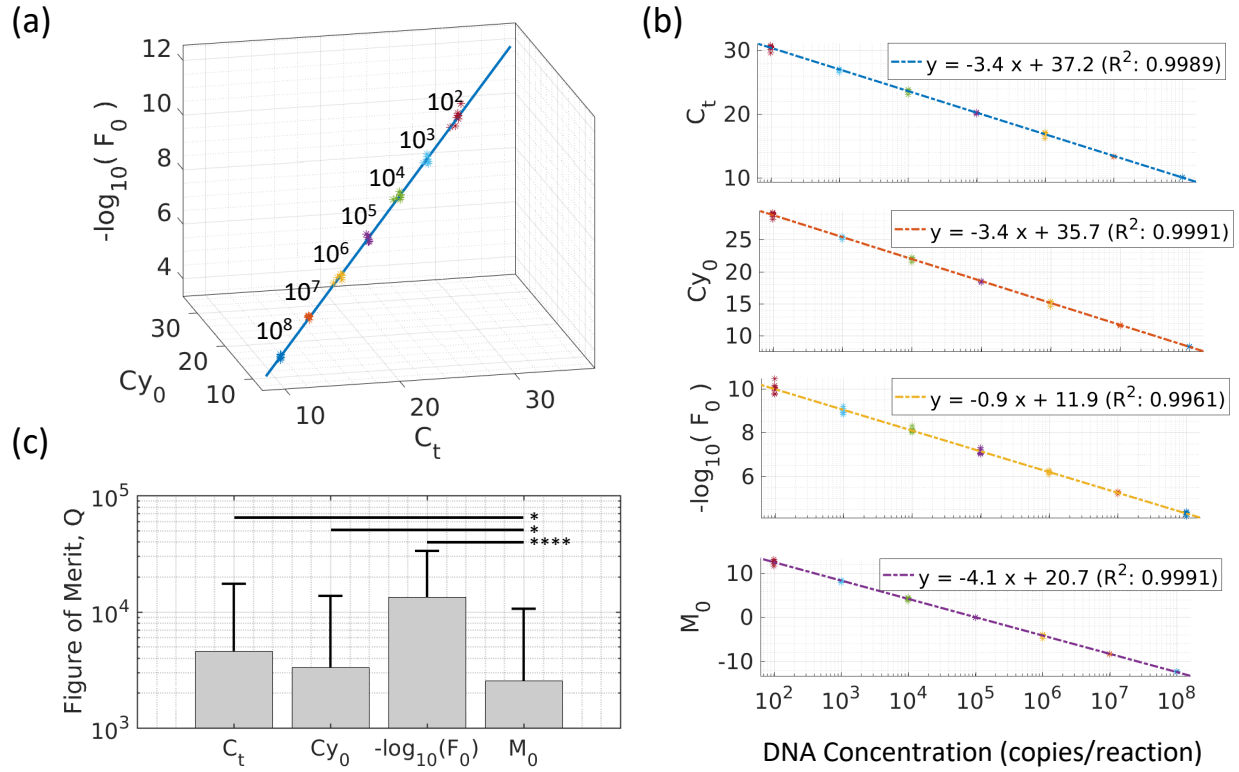


Figure 2: Evaluating quantification through the multidimensional standard curve and single-feature methods. (a) A multidimensional standard curve is constructed using  $C_t$ ,  $Cy_0$  and  $-\log_{10}(F_0)$  for lambda DNA with concentration values ranging from  $10^2$  to  $10^8$  (top right to bottom left). (b) The constructed standard curves using single-feature methods along with  $M_0$ . (c) The average figure of merit combining accuracy, precision and overall predictive power for all methods. A paired two-sided Wilcoxon signed rank test was performed between  $M_0$  and the other methods (\* $p$ -value  $< 0.05$  and \*\*\*\* $p$ -value  $< 0.0001$ ).

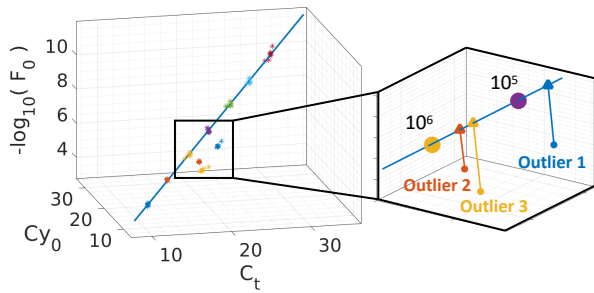


Figure 3: The multidimensional standard curve for lambda DNA along with three non-specific outliers. The right panel shows a zoomed region of the feature space with the mean of the replicates and the projection of the outliers onto the standard curve. The computed features/curve-fitting parameters of the three outliers and melting curve analysis are presented in Table S10 and Figure S1-S2 respectively.

outliers using a threshold. However, this similarity measure has two limitations: (i) There is an assumption that distances in different di-

rections are equally likely, which is intuitively untrue in the feature space because a change in one direction, e.g.  $C_t$ , does not impact the amplification curve as much as another, e.g.  $-\log_{10}(F_0)$ . (ii) There is no probabilistic measure that captures the distribution of the data and therefore the threshold for determining outliers must be chosen arbitrarily.

In order to tackle the two aforementioned limitations, the Mahalanobis distance,  $d$ , can be used. Clearly, by observing Fig 4 (b), the training data predominantly varies in a given direction. In order to visualise the Mahalanobis distance, the orthogonal view of the feature space (Fig 4 (b)) can be transformed into a new space (Fig 4 (c)) where the Euclidean distance is equivalent to the Mahalanobis distance in the original space. This is achieved by normalising the principal components of the training data.

The Mahalanobis distance from the multi-

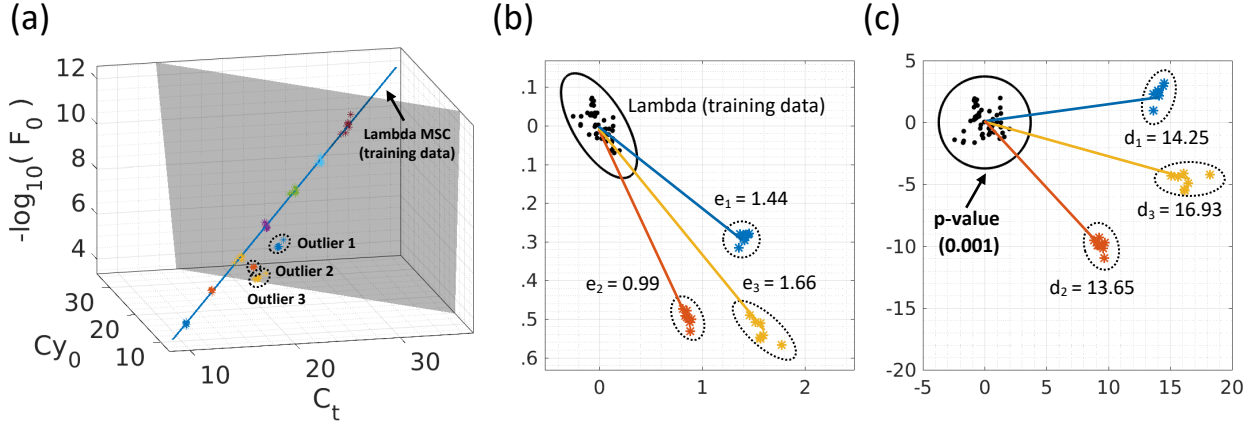


Figure 4: Multidimensional analysis using the feature space for detecting non-specific outliers. (a) MSC using  $C_t$ ,  $Cy_0$  and  $-\log_{10}(F_0)$  for lambda DNA along with three non-specific outliers. An arbitrary hyperplane orthogonal to the MSC is shown in grey. (b) Euclidean space: the view of the feature space when all the data points have been projected onto the aforementioned hyperplane. The Euclidean distance between the mean of the training data and the outliers ( $e_1$ ,  $e_2$  and  $e_3$ ). (c) Mahalanobis space: a transformed space where the Euclidean distance is equivalent to the Mahalanobis distance,  $d$ , in the Euclidean space. The black circle corresponds to a  $p$ -value of 0.001 using a  $\chi^2$ -distribution with 2 degrees of freedom.

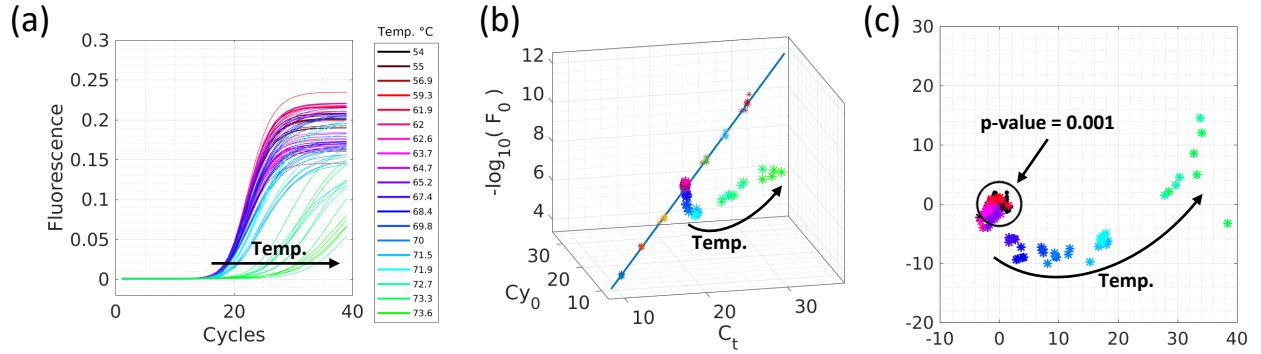


Figure 5: The effect of changing annealing temperature on detecting outliers using multidimensional standard curves. (a) Fluorescent amplification curves for lambda DNA ( $10^5$  copies per reaction) at temperatures ranging from  $54.0 - 73.6^\circ\text{C}$ . (b) The MSC constructed at  $62^\circ\text{C}$  using the features  $C_t$ ,  $Cy_0$  and  $-\log_{10}(F_0)$  along with data points obtained from the aforementioned fluorescent amplification curves. (c) The lambda standard and temperature variation data points in the Mahalanobis space. The black circle corresponds to a  $p$ -value of 0.001 using a  $\chi^2$ -distribution with 2 degrees of freedom.

mensional standard curve to the mean of the outliers is  $d_1 = 14.25$ ,  $d_2 = 13.65$  and  $d_3 = 16.93$ , respectively. In contrast with the Euclidean distances, it is observed that when considering the distribution of the data, the position of the outliers change. A useful property of  $d$  is that its squared value,  $d^2$ , follows a  $\chi^2$ -distribution if the data is approximately normally distributed. The hypothesis that the data is normally distributed is confirmed using the Henze-Zirkler test with a significance level of 0.05. Therefore, the distance can be converted

into a probability in order to determine if a data point is an outlier. Based on the  $\chi^2$ -distribution table with 2 degrees of freedom, any point further than 3.717 is 99.9% ( $p$ -value  $< 0.001$ ) likely to be an outlier. Since all the outliers have a Mahalanobis distance significantly greater than 3.717, they are confidently classified as outliers.

Aside from non-specific DNA amplification, another cause of outliers, especially in resource-limited settings, is due to inconsistent reaction efficiency's between the training and test data. In the following study, we use the annealing

temperature as a proxy for varying the efficiency of lambda DNA amplification. Fig 5 (a) shows the amplification curves for lambda DNA at  $10^5$  copies/reaction for temperatures ranging from  $54.0 - 73.6^\circ\text{C}$ . From observing the change in  $C_t$  it can be observed that, even though the product is specific (see Figure S3 for melting analysis), the quantification performance can be drastically affected. Current standard curve approaches have no heuristic measure to indicate whether any of these curves will be quantified poorly. However, when the data is viewed in the feature space (Fig 5 (b)) and the Mahalanobis space (Fig 5 (c)), it can be observed that when the amplification shape diverges from the curves belonging to the MSC (especially for low efficiencies), the Mahalanobis distance between the test data and the MSC increases. Therefore, this raises the question: can  $d$  be used to disregard *specific* outliers and therefore support quantification?

**Merging Quantification and Outlier Detection.** In order to investigate the use of the MSC and Mahalanobis distance for supporting quantification, we can compare the effect of removing outliers on the estimated quantification. Fig 6 (a) shows the average quantification as a function of the temperature for  $C_t$ ,  $Cy_0$ ,  $-\log_{10}(F_0)$  &  $M_0$  (bar plots) and visualises the temperatures at which the amplification curves are considered as outliers (shaded in red).

There are two key observations that can be made: (i) the quantification performance begins to deteriorate at temperatures above  $65.2^\circ\text{C}$  for all methods; (ii) amplification curves at temperatures above  $65.2^\circ\text{C}$  are considered as outliers based on a  $\chi^2$ -distribution with  $p < 0.001$ . These two observations coincide and, therefore, support the claim that the selected features extracted from the amplification curves contain sufficient information to disregard outliers and improve quantification performance. Fig 6 (b) shows the average relative error in estimated quantification for all the considered methods when using all data points and also disregarding outliers. It can be observed that quantification is improved by 59.9%, 53.9%, 93.9% and 44.6% for  $C_t$ ,  $Cy_0$ ,  $-\log_{10}(F_0)$  and  $M_0$ , respec-

tively. Notice that the benefits of multidimensional analysis using *all* of the features extends to enhancing quantification performance of *any* method, including single-feature methods.

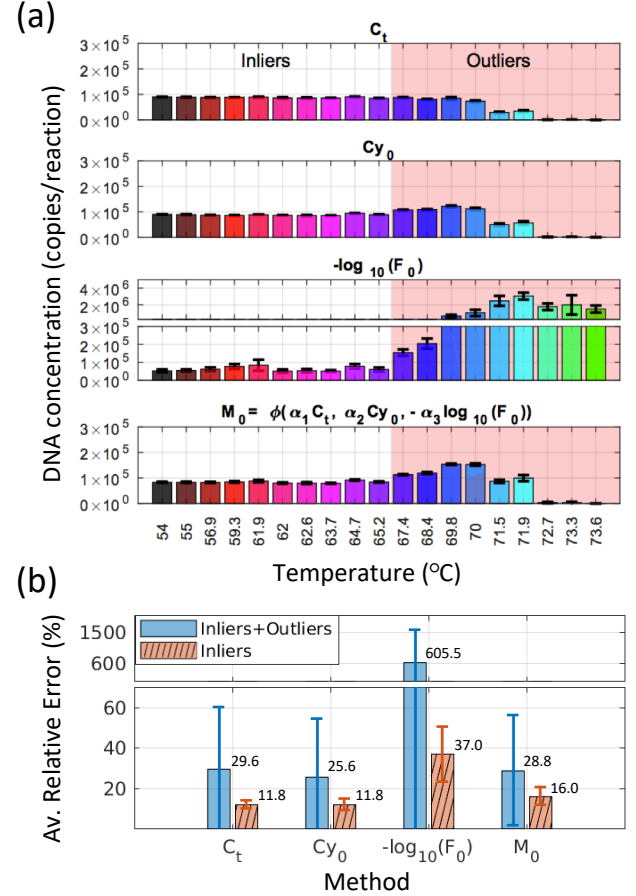


Figure 6: Merging quantification and outlier detection. (a) Average estimated quantification for lambda DNA at  $10^5$  copies per reaction for annealing temperatures ranging from  $54.0 - 73.6^\circ\text{C}$  using  $C_t$ ,  $Cy_0$ ,  $-\log_{10}(F_0)$  and  $M_0$ . The shaded region indicate outliers according to the MSC with  $p\text{-value} < 0.001$  based on a  $\chi^2$ -distribution with 2 degrees of freedom. Details for the quantification and outlier detection are provided in Table S11 and S12. (b) Average relative error for estimated quantification of lambda DNA at  $10^5$  copies/reaction across all annealing temperatures for every method. The solid bars represent the average relative error for all data points (including outliers and inliers) whereas the dashed bars only consider inliers. A paired two-sided Wilcoxon signed rank test was performed between  $M_0$  and the other methods with a confirmed significance ( $p\text{-value} < 0.0001$ ).

# Conclusion

Absolute quantification of nucleic acids in real-time PCR using standard curves is exceedingly important in several fields of biomedicine, although research has saturated in recent years. This is partially due to the simplicity of standard curves and the movement of research towards digital PCR (dPCR) because of the advantages it holds over qPCR, such removing the need for a standard curves. However, dPCR is currently not suitable for many applications given the cost and complexity of instruments. This paper presents a framework that shows the benefits of standard curves extend beyond absolute quantification when observed in a multidimensional environment. Consequently, this work opens the possibility for researchers from different fields to explore mathematical methods and applications that are enabled by the proposed framework.

The focus of current researchers is on the computation of a single value, referred to here as a *feature*, that is linearly related to template concentration. Therefore, there has been a gap in the literature in taking advantage of multiple features together. The potential reason for a lack of research in this area is because of the non-trivial benefits of combining linear features. The only intuitive interpretation of using several features is in the reliability of quantification. For example, instead of trusting a single feature, e.g.  $C_t$ , other features such as  $Cy_0$  and  $-\log_{10}(F_0)$  can be used to check if the quantification result is similar. This unidimensional way of thinking prevents several degrees of freedom and advantages that our proposed framework enables.

Three main capabilities are enabled by the framework proposed in this paper: (i) to optimise quantification performance based on a figure of merit; (ii) to detect outliers; and (iii) to measure how suitable an amplification curve is for the constructed MSC. The first capability provides a lower bound on the quantification performance of the framework to single best feature since this is a special case (e.g.  $M_0 = C_t$  when  $\alpha_1 = 1$ ,  $\alpha_2 = 0$  and  $\alpha_3 = 0$ ). The second and third capabilities are an application

of the MSC that was enabled through exploring the information gain captured by the elements of the feature space (e.g. Mahalanobis distance); which are typically meaningless or not considered in the unidimensional approach. In fact, applications of the MSC have already been developed. For example, in Rodriguez-Manzano et al., it was shown that multiple MSCs can be constructed in a shared feature space in order to simultaneously enhance quantification and multiplex 4 targets.<sup>16</sup>

The multidimensional approach is not completely unfamiliar in absolute quantification. The shape based outlier detection (SOD) takes a multidimensional approach in order to define a similarity measure between amplification curves.<sup>31</sup> However, there are two fundamental differences with the work of this paper. The first is that SOD relies on using a specific model for amplification, namely the 5-parameter sigmoid, and is therefore not a general framework. The second difference is that the pattern between the features in SOD and initial target concentration is unknown, therefore the SOD cannot be naturally integrated into the quantification process and is typically used as an add-on.<sup>32</sup> In other words, the multidimensional approach is only considered for outlier detection and quantification is still considered as unidimensional.

The contribution of this work can be accredited to the framework as a whole and the feature space which incorporates the multidimensional standard curve. Currently, the framework is limited to considering features that are linearly related to initial target concentration. This limitation is in fact a design choice given that there is a lack of other types of features available in the literature with non-linear relationships, and in order to reduce the complexity of the analysis. The second limitation is related to the feature space. The question arises as to whether sufficient information is captured between amplification curves in order to distinguish them in the feature space. For example, if two unrelated PCR reactions exhibit a perfectly symmetric sigmoidal amplification curve, their position in the feature space may potentially overlap. This limitation can be tackled from

a molecular perspective by tuning the chemistry in order to sufficiently change amplification curves without compromising the performance of the reaction (e.g. speed, sensitivity, specificity, etc).

In terms of future directions, there are many research paths that can be explored. Both the theory of the framework and its applications can be investigated. The results presented in this paper raise a number of questions: Can the proposed framework be used for emerging isothermal amplification chemistries? Is there any benefit of using more than 3 features? How many MSCs can the feature space accommodate for multiplex assays? How could the framework accommodate features with a non-linear relationship to initial template concentrations?

In conclusion, this paper presents a framework, multidimensional standard curve and the feature space - which presents many opportunities for researchers to explore new techniques and ideas. This methodology will also have huge potential for emerging diagnostic technologies with high-throughput such as ISFET arrays, where each reaction can have thousands of amplification curves and detecting outliers manually is infeasible.<sup>33–35</sup> We hope that by sharing these concepts, others will be able to adapt and enhance this work to meet their objectives and advance the field of nucleic acid research.

**Acknowledgement** We would like to acknowledge Imperial Confidence in Concepts - Joint Translational Fund, Wellcome Trust ISSF (PS3111EESA to P.G. and J.R.M.), EPSRC Pathways to Impact (PSE394EESA to P.G. and J.R.M.), EPSRC Global Challenge Research Fund (EP/P510798/1 to P.G. and J.R.M.) and EPSRC DTP (EP/N509486/1 to A.M.) for supporting this work. The research was also partially funded by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infection and Antimicrobial Resistance at Imperial College London in partnership with Public Health England (PHE; HPRU-2012-10047 to A.H.). The views expressed are those of the au-

thor(s) and not necessarily those of the NHS, the NIHR, the Department of Health or Public Health England.

## Supporting Information Available

The following files are available free of charge.

- Synthetic DNA sequences, numerical values of extracted features and sigmoidal fittings for lambda DNA standard, breakdown for Figure of Merit of  $C_t$ ,  $C_{y0}$ ,  $-\log_{10}(F_0)$  and  $M_0$ , numerical values of extracted features and sigmoidal fittings for non-specific outliers, numerical values of extracted features and sigmoidal fittings for temperature variation experiment, estimated quantification for temperature variation experiment, melting curve analysis for lambda DNA standard experiment, melting curve analysis for non-specific outlier detection experiment and melting curve analysis for temperature variation experiment.

## References

- (1) Higuchi, R.; Fockler, C.; Dollinger, G.; Watson, R. Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. *Nature Biotechnology* **1993**, *11*, 1026.
- (2) Heid, C. A.; Stevens, J.; Livak, K. J.; Williams, P. M. Real time quantitative PCR. *Genome research* **1996**, *6*, 986–994.
- (3) Gingeras, T. R.; Higuchi, R.; Kricka, L. J.; Lo, Y. D.; Wittwer, C. T. Fifty years of molecular (DNA/RNA) diagnostics. *Clinical chemistry* **2005**, *51*, 661–671.
- (4) Mackay, I. M.; Arden, K. E.; Nitsche, A. Real-time PCR in virology. *Nucleic acids research* **2002**, *30*, 1292–1305.
- (5) Bustin, S. A. Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays.

- Journal of molecular endocrinology* **2000**, *25*, 169–193.
- (6) Nolan, T.; Hands, R. E.; Bustin, S. A. Quantification of mRNA using real-time RT-PCR. *Nature protocols* **2006**, *1*, 1559.
  - (7) Girones, R.; Ferrús, M. A.; Alonso, J. L.; Rodriguez-Manzano, J.; Calgua, B.; de Abreu Corrêa, A.; Hundesa, A.; Carratala, A.; Bofill-Mas, S. Molecular detection of pathogens in water—the pros and cons of molecular techniques. *Water research* **2010**, *44*, 4325–4339.
  - (8) Caliendo, A. M. et al. Better Tests, Better Care: Improved Diagnostics for Infectious Diseases. *Clinical Infectious Diseases* **2013**, *57*, S139–S170.
  - (9) Ghani, A. C.; Burgess, D. H.; Reynolds, A.; Rousseau, C. Expanding the role of diagnostic and prognostic tools for infectious diseases in resource-poor settings. *Nature* **2015**, *528*, S50–S52.
  - (10) Misyura, M.; Sukhai, M. A.; Kulasigan, V.; Zhang, T.; Kamel-Reid, S.; Stockley, T. L. Improving validation methods for molecular diagnostics: application of Bland-Altman, Deming and simple linear regression analyses in assay comparison and evaluation for next-generation sequencing. *Journal of clinical pathology* **2018**, *71*, 117–124.
  - (11) Wittwer, C. T.; Herrmann, M. G.; Moss, A. A.; Rasmussen, R. P. Continuous fluorescence monitoring of rapid cycle DNA amplification. *Biotechniques* **1997**, *22*, 130–139.
  - (12) Wittwer, C.; Ririe, K.; Rasmussen, R. Fluorescence monitoring of rapid cycle PCR for quantification. **1998**, 129–144.
  - (13) Freeman, W. M.; Walker, S. J.; Vrana, K. E. Quantitative RT-PCR: pitfalls and potential. *Biotechniques* **1999**, *26*, 112–125.
  - (14) Guescini, M.; Sisti, D.; Rocchi, M. B.; Stocchi, L.; Stocchi, V. A new real-time PCR method to overcome significant quantitative inaccuracy due to slight amplification inhibition. *BMC bioinformatics* **2008**, *9*, 326.
  - (15) Rutledge, R. Sigmoidal curve-fitting redefines quantitative real-time PCR with the prospective of developing automated high-throughput applications. *Nucleic acids research* **2004**, *32*, e178–e178.
  - (16) Rodriguez-Manzano, J.; Moniri, A.; Malpartida-Cardenas, K.; Dronavalli, J.; Davies, F.; Holmes, A.; Georgiou, P. Simultaneous Single-Channel Multiplexing and Quantification of Carbapenem-Resistant Genes Using Multidimensional Standard Curves. *Analytical Chemistry* **2019**, *91*, 2013–2020.
  - (17) Van Der Maaten, L.; Postma, E.; Van den Herik, J. Dimensionality reduction: a comparative. *J Mach Learn Res* **2009**, *10*, 66–71.
  - (18) Rao, X.; Lai, D.; Huang, X. A new method for quantitative real-time polymerase chain reaction data analysis. *Journal of Computational Biology* **2013**, *20*, 703–711.
  - (19) Coleman, T. F.; Li, Y. An interior trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on optimization* **1996**, *6*, 418–445.
  - (20) Coleman T Li, Y. On the Convergence of Reflective Newton Methods for Large-Scale Nonlinear Minimization Subject to Bounds. *Mathematical Programming* **1994**, *67*, 189224.
  - (21) Levenberg, K. A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics* **1944**, *2*, 164–168.
  - (22) Marquardt, D. W. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial*



- and *Applied Mathematics* **1963**, 11, 431–441.
- (23) Friedman, J.; Hastie, T.; Tibshirani, R. *The elements of statistical learning*; Springer series in statistics New York, NY, USA:, 2001; Vol. 1.
  - (24) Fischler, M. A.; Bolles, R. C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM* **1981**, 24, 381–395.
  - (25) De Maesschalck, R.; Jouan-Rimbaud, D.; Massart, D. L. The mahalanobis distance. *Chemometrics and intelligent laboratory systems* **2000**, 50, 1–18.
  - (26) Coomans, D.; Broeckaert, I.; Derde, M.; Tassin, A.; Massart, D.; Wold, S. Use of a microcomputer for the definition of multivariate confidence regions in medical diagnosis based on clinical laboratory profiles. *Computers and biomedical research* **1984**, 17, 1–14.
  - (27) Nelder, J. A.; Mead, R. A simplex method for function minimization. *The computer journal* **1965**, 7, 308–313.
  - (28) Lagarias, J. C.; Reeds, J. A.; Wright, M. H.; Wright, P. E. Convergence properties of the Nelder–Mead simplex method in low dimensions. *SIAM Journal on optimization* **1998**, 9, 112–147.
  - (29) Thode, H. C. Testing for normality. **2002**, 164.
  - (30) Monteiro, J.; Widen, R. H.; Pignatari, A. C.; Kubasek, C.; Silbert, S. Rapid detection of carbapenemase genes by multiplex real-time PCR. *Journal of Antimicrobial Chemotherapy* **2012**, 67, 906–909.
  - (31) Sisti, D.; Guescini, M.; Rocchi, M. B.; Tibollo, P.; D’Atri, M.; Stocchi, V. Shape based kinetic outlier detection in real-time PCR. *BMC bioinformatics* **2010**, 11, 186.
  - (32) Guescini, M.; Sisti, D.; Rocchi, M. B.; Panebianco, R.; Tibollo, P.; Stocchi, V. Accurate and precise DNA quantification in the presence of different amplification efficiencies using an improved Cy0 method. *PloS one* **2013**, 8, e68481.
  - (33) Rodriguez-Manzano, J.; Chia, P. Y.; Yeo, T. W.; Holmes, A.; Georgiou, P.; Yacoub, S. Improving Dengue Diagnostics and Management Through Innovative Technology. *Current infectious disease reports* **2018**, 20, 25.
  - (34) Miscourides, N.; Yu, L.-S.; Rodriguez-Manzano, J.; Georgiou, P. A 12.8 k current-mode velocity-saturation ISFET array for on-chip real-time DNA detection. *IEEE transactions on biomedical circuits and systems* **2018**, 1–13.
  - (35) Moser, N.; Rodriguez-Manzano, J.; Lande, T. S.; Georgiou, P. A scalable ISFET sensing and memory array with sensor auto-calibration for on-chip real-time DNA detection. *IEEE transactions on biomedical circuits and systems* **2018**, 12, 390–401.

# Graphical TOC Entry

